DOCUMENT RESUME

ED 346 161 TM 018 547

AUTHOR Vispoel, Walter P.; And Others
TITLE How Review Options and Administration Modes Influence Scores on Computerized Vocabulary Tests.
PUB DATE Apr 92
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adaptive Testing; *College Students; Comparative Testing; *Computer Assisted Testing; Concurrent Validity; Difficulty Level; Error Correction; Feedback; Higher Education; *Review (Reexamination); *Scores; Test Anxiety; Test Reliability; Test Wiseness; *Vocabulary Skills
IDENTIFIERS *Fixed Item Tests

ABSTRACT

The effects of review options (the opportunity for examinees to review and change answers) on the magnitude, reliability, efficiency, and concurrent validity of scores obtained from three types of computerized vocabulary tests (fixed item, adaptive, and self-adapted) were studied. Subjects were 97 college students at a large midwestern university who each completed one of the vocabulary tests and several measures of attitudes about review, item difficulty, and test anxiety. Review modestly enhanced test performance, slightly decreased measurement precision, moderately increased total testing time, affected concurrent validity, and was strongly favored by examinees. Computerized tests do not necessarily yield equivalent results, and such tests may have to be equated to ensure fair use of test scores. Differences in performance favoring paper-and-pencil tests in some prior studies occurred because review options were excluded from the computerized tests. Results for administration mode are inconclusive. Compared to the other types, the fixed-item test yielded the least desirable results because scores were lowest, least reliable, and most susceptible to test anxiety effects. The choice between self-adapted and adaptive tests seems to depend on examinee anxiety level. Item difficulty suggestion, rather than answer feedback, is the predominant factor facilitating performance on self-adapted tests. Included are 3 tables, 4 graphs, and 38 references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
WALTER P. VISPOEL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346161

How Review Options and Administration Modes Influence Scores on Computerized Vocabulary Tests

Walter P. Vispoel, Tianyou Wang, Roberto de la Torre, Tim Bleiler, & Jonathan Dings

The University of Iowa

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1992

TM018573

BEST COPY AVAILABLE

The vast majority of paper and pencil teacher-constructed and standardized tests developed over the last 75 years have provided examinees with the opportunity for reviewing and changing their answers. These options, however, are not always provided on computerized tests, which increasingly are being used as alternatives to paper and pencil tests. The primary goal of the present study was to examine the effects of review options on the magnitude, reliability, efficiency, and concurrent validity of scores obtained from three types of computerized vocabulary tests -- fixed-item, adaptive, and self-adapted.

Background

The effects of reviewing and changing answers on fixed-item paper and pencil tests have been studied since the late 1920s. Nearly all studies have shown that reviewing answers modestly improves test performance (cf., Crawford, 1928; Crocker & Benson, 1980; Cummings, 1981; Matthews, 1929; McMorris & Weideman, 1986; McMorris, DeMers, & Schwarz, 1987; Mueller & Wasser, 1977; Smith, White, & Coop, 1979). The effects of review on computerized tests, however, has received very limited attention. Most relevant research in this area has involved studies comparing paper and pencil and computerized versions of fixed-item tests. These studies typically have shown that test performance is higher on the paper and pencil tests (see Mazzeo & Harvey, 1988 and Bunderson, Inouye, and Olsen, 1989 for reviews of these studies), and the absence of review options on the computerized tests often is cited as a possible reason for these differences.

Differential performance on tests with and without review options raises important questions about fairness and comparability of test results and about the emotional climate of test administration settings. Individuals who are not allowed to review their answers may receive systematically lower scores than individuals who are allowed to review. Excluding review opportunities on computerized tests also can be upsetting and frustrating to examinees because they are so accustomed to having these options on paper and pencil tests (Buhr & Legg, 1989; de la Torre & Vispoel, 1991; Moe & Johnson, 1986; Baghi, Gabrys, & Ferrara, 1991; Gershon & Bergstrom, 1991). These emotional reactions, in turn, may further inhibit examinees from performing at their best.

Issues of equity and negative affect are particularly relevant when using computerized adaptive and self-adapted tests because review options generally have been excluded from these tests. In an adaptive test, examinees receive items targeted to their ability levels and avoid items that are either too difficult or too easy. A common strategy is to begin with an item of average difficulty, administer an easier item following an incorrect response or a harder item following a correct response, and end the test when a fixed number of items is administered or when a desired level of reliability is reached. Research has shown that adaptive tests provide greater measurement precision than fixed-item tests when test length is held constant and generally require about half as many items to match a fixed-item test's reliability (Brown & Weiss, 1977; Lord, 1977; McBride, 1986; McBride

and Martin, 1983; Moreno, Wetzel, McBride & Weiss, 1984; Stocking, 1987: Sympson, Weiss, & Ree, 1982; Urry, 1977; Vispoel, in press; Vispoel and Coffman, 1992; Weiss, 1982). Review options typically have been excluded from adaptive tests because such options may reduce these gains in measurement precision and efficiency. Measurement precision will change if the administered items provide less information about the examinee's post-review ability estimate than they did about the examinee's pre-review ability estimate. Test efficiency will be reduced because examinees will spend additional time reviewing items.

Only one study to date has examined the effects of review on scores from computerized adaptive tests. In the part of that study most relevant to the present investigation, Lunz, Bergstrom, and Wright (1991) compared the performance of 220 examinees before and after they reviewed answers to a practice version of a medical certification test. Each examinee responded to a minimum of 50 and maximum of 240 items ($\underline{M}$ = 96). The test was terminated after a reliable pass/fail decision was reached and at least 50 items were administered. Results indicated that (a) 2% of item answers were altered following review, (b) mean test performance improved slightly following review, (c) 99% of pass/fail decisions were equivalent before and after review, (d) 1% of the examinees who failed the test before review passed it after review, and (e) relative efficiency as indexed by the test information ratio decreased by 1% following review. Since the results of this study indicated that performance improved slightly with little change in reliability and relative efficiency, the authors concluded that review options should be included on such tests.

No prior study had examined the effects of review on computerized self-adapted tests. A self-adapted test is similar to an adaptive test except that the examinee rather than the computer chooses the desired difficulty level of each item before responding. When compared to computerized fixed-item and adaptive testing strategies, self-adapted tests have yielded higher scores or scores less influenced by test anxiety (O'Donnell and Rocklin, 1991; Rocklin and O'Donnell, 1987, 1991; Vispoel, 1991; Wise, Plake, Johnson, & Roos, 1991). These findings have been attributed to the opportunity of examinees to adjust the difficulty of test items to match their motivational and affective states throughout the test. The affective benefits of self-adapted tests also have been supported by examinees' strong endorsement of item difficulty selection options after they have completed self-adapted tests (Rocklin & O'Donnell, 1991; Vispoel, 1991; Vispoel & Rocklin, 1992). The primary drawbacks to self-adapted tests are that they are usually less reliable than adaptive tests when test length is held constant and they require additional administration time to allow examinees to choose item difficulty levels (O'Donnell and Rocklin, 1991; Rocklin and O'Donnell, 1991; Vispoel, 1991; Vispoel & Rocklin, 1992; Wise, Plake, Johnson, & Roos, 1991).

Review options have been inappropriate for the self-adapted tests used in prior studies because feedback was provided after each item informing examinees about whether their responses were correct or incorrect. Eliminating such feedback, however, permits the examination of review

effects on these tests and removes the confounding of feedback and item difficulty selection that has complicated the interpretation of results from prior studies.

Purposes and Significance of the Present Study

The primary purpose of the present study was to examine the effects of review options and administration modes on the magnitude, reliability, efficiency, and concurrent validity of scores yielded by computerized vocabulary tests. A secondary purpose was to examine examinee's attitudes about the importance of including review and item difficulty selection options on computerized tests. This study extended prior research in several ways: (a) the effects of review were examined simultaneously for three computerized administration modes (fixed-item, adaptive, self-adapted) and for individuals at varying levels of test anxiety, (b) the effects of item difficulty selection in the self-adapted tests were examined independent of answer feedback, and (c) the concurrent validity of scores was compared.

Based on the related research, the following predictions were made.

1. Review would improve performance modestly on all tests.
2. Performance for individuals with high test anxiety would be higher on the self-adapted test than on the adaptive and fixed-item tests.
3. Review would decrease measurement precision slightly on the adaptive test.
4. Measurement precision would be higher on the adaptive test than on the fixed-item and self-adapted tests.
5. Review would increase administration time on all tests.
6. The self-adapted test would require more time than the other tests.
7. Examinees would favor the inclusion of review options on all tests.
8. Examinees who took the self-adapted test would favor item selection options more than examinees who took adaptive and fixed-item tests.

Method

Ninety-seven college students (24% male, 76% female; mean age = 21.68; 10% freshman, 39% sophomores, 26% juniors, 13% seniors, 10% graduate) from a large midwestern university participated in the study as part of a research option in an introductory educational psychology course. During the 50-60 minute research session, participants completed the following measures in this order: (a) an investigator-designed questionnaire assessing demographic information and self-reported grade-point-average, (b) the verbal self-concept subscale from the Self Description Questionnaire III (Marsh, 1987), (c) the Test Anxiety Inventory (TAI; Spielberger, 1980), (d) a computerized vocabulary test, (e) an investigator-designed questionnaire assessing the importance of including review and item difficulty

selection options on computerized tests, and (f) a 60-item paper-and-pencil vocabulary test consisting of retired or previously piloted items from the Iowa Tests of Educational Development (Feldt, Forsyth, & Alnot, 1986). Self-reported grade-point-average, verbal self-concept scores, and the paper-and-pencil ITED scores served as criterion measures in the concurrent validity analysis. Alpha coefficients for the paper-and-pencil vocabulary test, self-concept scale, and Test Anxiety Inventory were .87, .91, and .95, respectively.

Participants were randomly assigned to three test administration conditions: fixed-item (n=32), adaptive (n=33), self-adapted (n=32). The computerized adaptive and self-adapted tests used a pool of the 200 items from five equated forms of the vocabulary subtest of the Iowa Tests of Educational Development (Feldt, Forsyth, & Alnot, 1986). This item pool is more than twice the size of any used in previous self-adapted testing studies. The 5-option multiple-choice items from the pool were administered to over 10,000 subjects and calibrated using a modified three-parameter logistic model (i.e., c values were fixed at .15). All computerized tests were administered on IBM microcomputers and terminated at 30 items. The adaptive test employed EAP ability estimation (Bock & Mislevy, 1982) and maximum information item selection procedures. Examinees chose items from eight difficulty levels when taking the self-adapted test. The fixed-item test consisted of the last 30 items from one of the ITED vocabulary test forms included in the adaptive test item pool. Pilot data indicated that these items were well matched to the ability levels of previous students enrolled in the educational psychology course. The tests were administered using CATSYS (de la Torre & Vispoel, 1991) -- a new adaptive testing computer package shown to reproduce true ability scores with measurement accuracy and efficiency comparable or superior to that of MicroCAT (Assessment Systems Corporation, 1989). Examinees in all conditions were instructed to provide the best answer to each question the first time through the test and that they would be able to review and change their answers later. Examinees also were told that their scores would be rank-ordered, graded and posted according to their student IDs at the completion of the experiment and that individuals with the five highest scores each would receive $20. A 20-minute time limit was established for each test. These steps were taken to motivate the examinees to do their best. Responses, ability scores, and their variances were stored in the computer after each item was administered and reviewed.

Results

Data Analyses

The effects of review, administration mode, and anxiety on ability scores, measurement precision, and testing time were evaluated in a series of split-plot ANOVAs. Each analysis included one within-subject factor -- review (before and after), and two between-subjects factors -- anxiety (with TAI total scores entered as a continuous variable) and administration mode (fixed-item, adaptive, self-adapted). A hierarchical regression approach was used. In the within-subject analyses, review was

entered first followed with the review by anxiety, review by testing mode, and review by anxiety by testing mode interactions. In the between-subjects analyses, test anxiety, testing mode, and the testing mode by anxiety interaction were entered in that order. Separate between-subjects analyses also were conducted for ability scores, measurement precision, and testing time before and after review and for the attitudinal indices. Concurrent validity analyses were based on comparisons of zero-order correlations. An alpha level of .05 was adopted for all analyses. The ANOVA results are provided in Table 1. The major findings from these analyses are described below.

Test Performance

As can be seen from the ANOVA results for test performance in Table 1, the review effect was significant in the within-subject analysis and the anxiety, testing mode, and anxiety by testing mode effects were significant in the between-subjects analyses. Test performance, as predicted, was higher following review and the gain was modest, representing an average increase of .043 on the theta scale. The descriptive statistics for answer changing behavior, provided in Table 2, show that gains (wrong to right changes) outnumbered losses (right to wrong changes) for all testing modes. On average, examinees' number correct scores increased by .456 points (or 1.5% correct) following review. The overall gain to loss ratio was 2.10 to 1 in terms of total items changed and 3.75 to 1 in terms of examinees whose scores changed following review. Answers to 5.67% of all items were changed after review, and 58.26% of the examinees changed responses to at least one item.

The anxiety by testing mode interactions, which take precedence over the anxiety and testing mode main effects, are depicted in Figure 1. The pattern of results is essentially the same both before and after review. As expected, performance on the self-adapted test was reasonably consistent across anxiety levels but highly test-anxious examinees performed noticeably lower on the fixed-item and adaptive tests. The debilitative effects of test anxiety were especially pronounced on the fixed-item test. Overall performance on the fixed-item test also was lower than performance on the other tests.

Measurement Precision

As Table 1 shows, significant review and testing mode main effects were found for measurement precision. Consistent with Lunz et al. (1991), error variance increased significantly following review, but the increase ($\underline{M}$ = .0020) was small and of little practical significance. The total group test information ratio (precision after review over precision before review) was .97, which reflects a loss of 3% in relative efficiency. Error variance, as expected, increased in the adaptive testing condition ($\underline{M}$ = 0016). Similar increases in error, although not predicted, were found for the fixed-item test ($\underline{M}$ = .0016) and self-adapted test.($\underline{M}$ = .0028).

The testing mode main effects for measurement precision, which were significant both before and after review, revealed a similar trend. As expected, error variance was lower for the adaptive test ($\underline{M}$ = .0428, before review; $\underline{M}$ = .0444, after review) than for the self-adapted test ($\underline{M}$ = 0600, before review;

M = .0628, after review) and the fixed-item test (M = .1069, before review; M = .1085, after review). Tukey pairwise comparison tests indicated that the differences between the fixed- item and other tests were statistically significant but differences between the adaptive and self-adapted tests were not.

Testing Time

The within-subject ANOVA results for testing time in Table 1 indicate that the review main effect and review by testing mode interaction were significant. Overall, examinees spent more time on the first pass through the tests (M = 7 min. and 20 sec.) than on reviewing them (M = 2 minutes and 44 sec.). The review by testing mode interaction, depicted in Figure 2, shows that the self-adapted test required more time before review (M = 8 min. and 34 sec.) than did the adaptive test (M = 6 min. and 40 sec.) and fixed-item test (M = 6 min. and 47 sec.). The additional time taken on the self-adapted tests before review was expected because examinees must choose item difficulty levels in addition to answering the items. The time spent reviewing answers, however, was similar across testing modes. On average (and excluding directions), the self-adapted, adaptive, and fixed item tests required 11 min. 50 sec., 9 min. 44 sec., and 9 min. 53 sec. of total testing time, respectively. The inclusion of review options increased testing time across tests by about 27%. The inclusion of both review and item difficulty selection options increased the testing time on the self-adapted test by about 43%.

The between-subject ANOVA results for testing time in Table 1 indicate that testing mode effects were significant before review but not during review, confirming the interpretation of the review by testing mode interaction depicted in Figure 2. Significant main effects for test anxiety also were found in the between-subjects ANOVAs both before and during review. The anxiety effects, depicted in Figure 3, indicate that highly test-anxious examinees spent more time both responding to and reviewing items than did less test -anxious examinees.

Concurrent validity

As Table 3 shows, differences in validity coefficients before and after review were negligible in all cases. These results are not surprising given that the correlations between ability scores before and after review were .9886, .9847, .9847, and .9812 for the fixed-item, adaptive, self-adapted, and total groups, respectively. Three significant differences in validity coefficients was found for testing modes. The correlation between ability (theta) scores and self-reported GPA for the adaptive test was significantly higher than the corresponding correlation for the fixed-item test ($Z = 1.98$, $p < .05$), and the correlation between ability scores and verbal self-concept for the adaptive test was significantly higher than the corresponding correlation for the self-adapted test both before ($Z = 2.06$, $p < .05$) and after review ($Z = 2.01$, $p < .05$). These differences in validity coefficients for verbal self-concept, however, may indicate that verbal self-concept is more suitable as an individual difference variable than as a criterion measure. As was the case for test anxiety, self-adapted testing may reduce the influence of verbal self-concept on vocabulary test scores.

Attitudes about including review and item difficulty selection options

The ANOVA results for attitudes, provided in Table 1, reveal no significant effects for the perceived importance of including review options. As expected, however, examinees across testing modes strongly favored the inclusion of such options ($\underline{M}$ = 6.62 on an 8-point scale, with 87% favoring inclusion).

A significant testing mode main effect was found for the perceived importance of including item difficulty selection options. Consistent with expectations, examinees who took the self-adapted test ($\underline{M}$ = 7.16) endorsed the inclusion of item difficulty selection options much more than individuals who took the adaptive ($\underline{M}$ = 4.14) or fixed-item ($\underline{M}$ = 4. 59) tests.

Summary and Discussion

Permitting review on the computerized vocabulary tests modestly enhanced test performance, slightly decreased measurement precision, moderately increased total testing time, had little effect on concurrent validity, and was strongly favored by examinees. This combination of findings, although generally supportive of review options, provides no definitive answer as to whether these options should be included or excluded on computerized tests. It could be argued that review opportunities should be provided because test performance may improve and examinees clearly desire and use these options. Alternatively, it could be argued that review should be excluded due to losses in test efficiency and the high correlation between scores before and after review. What is clear, however, is that computerized tests with and without review do not necessarily yield equivalent results and that such tests may have to equated to ensure fair use of test scores. The present results also support the premise that the differences in performance favoring paper and pencil tests over computerized tests reported in some prior studies occurred because review options were excluded on the computerized tests.

As was the case with review options, the results for administration mode do not provide a simple answer as to which procedure is best because each mode had its strengths and weaknesses. Among the testing modes investigated, the fixed-item test yielded the least desirable results because scores were lowest, least reliable, and most susceptible to test anxiety effects. The choice between self-adapted and adaptive tests seems to depend on the examinee's anxiety level. Replicating findings from Rocklin and O'Donnell, (1991) and O'Donnell and Rocklin (1991), the present self-adapted test provided more construct-valid scores for individuals with high test anxiety, and would appear to be the most viable assessment procedure for these individuals. Differences in performance among tests, however, were less marked for individuals with low test anxiety, and the efficiency provided by the adaptive test would make it a more desirable alternative for those individuals. The differences in performance among testing modes, like those found for review, imply that scores derived from fixed-item, adaptive, and self-adapted tests may need to be equated to ensure fair use of test results.

In addition to replicating several prior findings about administration mode effects, the present study provided at least three new findings. One finding was that the facilitative effects of the self-adapted test could be unambiguously attributed to item difficulty selection options because answer feedback was not provided. This finding coupled with that from Wise, Plake, Johnson, and Roos (1991), in which answer feedback was provided on both the adaptive and self-adapted tests, suggests that item difficulty selection rather than answer feedback is the predominant factor facilitating performance on self-adapted tests.

A second new finding was that the adaptive and self-adapted tests yielded comparable measurement precision at a length of 30 items. This result occurred in part because the most discriminating item available at each selected difficulty level always was administered. Although this strategy was effective in producing reliable scores, the measurement precision for self-adapted tests may be enhanced even further by choosing the item within the selected difficulty category that provides the most information about the examinee's most current ability estimate. The relative effectiveness of these and other item selection algorithms for self-adapted tests is an important topic for future research.

A final and unanticipated finding was that performance was higher on the adaptive test than on the fixed-item test. This may have occurred because items on the adaptive test, which explicitly were targeted to each examinee's ability level, optimally challenged examinees without boring or discouraging them. The motivational properties of adaptive testing algorithms is another area of research that merits future investigation.

Several additional suggestions can be made for extending the present study. First, the effects of review options and computerized administration modes should be investigated using larger samples of examinees, different test termination criteria, different test content, and more realistic classroom and standardized testing situations. Second and as recommended by McMorris et al (1987), examinees should be interviewed to determine why they altered their responses to each item so that effective and ineffective answer changing strategies can be identified. Third, other individual difference variables (e.g., impulsivity, risk-taking, cognitive structure, self-concept, self-efficacy, prior ability) should be examined to determine how they interact with review and self-adapted testing behavior. Finally, the effects of item difficulty selection and feedback of self-adapted tests should be systematically manipulated to determine their individual and interactive effects on test performance.

References

Assessment Systems Corporation (1989). User's manual for MicroCAT. St. Paul, MN: Author.

Baghi, H., Gabrys, R., & Ferrara, S. (1991, April). Applications of computer-adaptive testing in Maryland. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 37, 431-444.

Brown, J. M., & Weiss, D. J. (1977). An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.

Buhr, D. C., & Legg, S. M. (1989, March). Investigating the validity of a computerized adaptive test for different examinee groups. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989) The four generations of computerized educational measurement. In R. L. Linn (Ed.) Educational measurement (3rd. edition). New York: Macmillan.

Crawford, C. (1928). The technique of study. Boston: Houghton Mifflin.

Crocker, L., & Benson. J. (1980). Does answer-changing affect test quality? Measurement and Evaluation in Guidance, 12, 223-239.

Cummings, O. W. (1981). Impact of response changes on objective test characteristics and outcomes for junior high school students. Measurement and Evaluation in Guidance, 14, 32-37.

de la Torre, R., & Vispoel, W. P. (1991, April). The development and evaluation of a computerized adaptive testing system. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Feldt, L S., Forsyth, R. A., & Alnot, S. D. (1986). Iowa Tests of Educational Development. Iowa City: The University of Iowa.

Gershon, R. C., & Bergstrom, B. (1991, April). Individual differences in computer adaptive testing: Anxiety, computer literacy and satisfaction. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1991, April). The effect of review on student ability and test efficiency for computer adaptive tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Marsh, H. W. (1987). The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. Journal of Educational Measurement, 24, 17-39.

Matthews, C. O. (1929). Erroneous first impressions on objective tests. Journal of Educational Psychology, 20, 280-286.

BEST COPY AVAILABLE

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

McBride, J. R. (1986). A computerized adaptive edition of the Differential Aptitude Tests. (ERIC Document Reproduction Service No. ED 285 918). Psychological Corporation.

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

McMorris, R. F., & Weiderman, A. H. (1986). Answer changing after instruction on answer changing. Measurement and Evaluation in Counseling and Development, 18, 93-101.

McMorris, R. F., DeMers, L. P., & Schwarz, S. P. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. Journal of Educational Educational Measurement, 24, 131-143.

Moe, K. C., & Johnson, M. F. (1986). Participants' reactions to computerized testing. ERIC Document Reproduction Service No. ED 282935.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. Applied Psychological Measurement, 8(2), 155-163.

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. Journal of Educational Measurement, 14, 9-13.

O'Donnell, A. M., & Rocklin, T. R., (1991). The effects of examinee choice of item difficulty on test performance Manuscript submitted for publication .

Rocklin, T. R., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. Journal of Educational Psychology, 79, 315-319.

Rocklin, T. R., & O'Donnell, A. M. (1991, April). An empirical investigation of self-adapted and maximum information item selection Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Smith, M., White, K., & Coop, R. (1979). The effect of item type on the consequences of changing answers on multiple choice tests. Journal of Educational Measurement, 16, 203-208.

Spielberger, C. D. (1980). Preliminary professional manual for the Test Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Stocking, M..L., (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 263-277.

Sympson, J. B., Weiss, D. J., & Ree, M. (1982). Predictive validity of conventional and adaptive tests in an Air Force training environment. (AF HRL-TR-81-40). Brooks Air Force Base. TX: USAF Human Resources Laboratory.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14(2), 181-196.

Vispoel, W. P. (1991, November). Self-adapted testing: A new method for assessing musical ability. Paper presented at the annual meeting of the Association for the Development of Computer-based Instructional Systems, St. Louis.

Vispoel, W. P. (in press). Improving the measurement of tonal memory with computerized adaptive tests. Psychomusicology.

Vispoel, W. P. & Coffman, D. D. (1992). Computerized adaptive testing of music-related skills. Bulletin of the Council for Research In Music Education, 112, 29-49.

Vispoel, W. P., & Rocklin, T. R. (1992). Fixed-item, adaptive, and self-adapted testing of vocabulary skills. Manuscript in progress.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement. 6(4), 473-492.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1991, April). A comparison of self-adaptive and algorithmic adaptive achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Table 1
ANOVA Results for Test Performance, Measurement Precision, Efficiency and Attitudes

EFFECTS		Theta	Error Variance	Testing Time	Importance of Review	Importance of Difficulty Selection
	df	F	F	F	F	F
FULL MODEL						
Review	1,91	9.76**	5.92*	457.94***		
Review x Anxiety	1,91	1.63	0.11	0.04		
Review x Test Mode	2,91	1.03	0.28	8.18***		
Review x Test Mode x Anxiety	2,91	0.63	0.97	0.84		
Anxiety	1,91	20.34***	0.06	7.41**	0.06	1.17
Test Mode	2,91	7.06**	16.46***	4.03*	0.99	19.89***
Test Mode x Anxiety	2,91	4.45*	0.18	0.08	0.35	1.71
BEFORE REVIEW						
Anxiety	1,91	19.75***	0.05	4.32*		
Test Mode	2,91	7.02**	17.37***	7.75***		
Test Mode x Anxiety	2,91	4.56*	0.20	0.32		
AFTER REVIEW						
Anxiety	1,91	20.53***	0.07	6.99**		
Test Mode	2,91	6.99**	15.41***	0.44		
Test Mode x Anxiety	2,91	4.27*	0.17	0.26		

* $p<.05$, ** $p<.01$, *** $p<.001$

TABLE 2
Descriptive Statistics for Test Performance and Answer Changing Behavior

Index	Fixed-Item	Adaptive	Self-Adapted	Total
Theta				
Before Review	1.056	1.579	1.573	1.399
After Review	1.083	1.653	1.601	1.442
Difference (After-Before)	0.027	0.073	0.028	0.043
Number Correct				
Before Review	21.727	22.281	21.781	21.928
After Review	21.970	23.000	22.188	22.382
Difference (After-Before)	0.243	0.719	0.407	0.454
Item Changes				
Average Number	1.24	1.53	2.75	1.83
Average Percent	3.76	4.79	8.59	5.67
Item Gains (W to R)				
Average Number	0.52	0.91	1.19	0.87
Average Percent	1.56	2.83	3.71	2.68
Item Losses (R to W)				
Average Number	0.27	0.19	0.78	0.41
Average Percent	0.83	0.59	2.44	1.28
Item (W to W)				
Average Number	0.45	0.44	0.78	0.56
Average Percent	1.37	1.37	2.44	1.72
Item Gain/Loss Ratio	1.89:1	4.83:1	1.52:1	2.10:1
Examinee Score Changes				
Average Number	13.00	19.00	25.00	57.00
Average Percent	39.39	59.38	78.13	58.76
Score Gains				
Average Number	9.00	18.00	18.00	45.00
Average Percent	27.27	56.25	56.25	46.39
Score Losses				
Average Number	4.00	1.00	7.00	12.00
Average Percent	12.12	3.13	21.88	12.37
Scores Unchanged				
Average Number	20.00	13.00	7.00	40.00
Average Percent	60.61	40.63	21.88	41.24
Examinee Gain/Loss Ratio	2.25:1	18.00:1	2.57:1	3.75:1

TABLE 3
Validity Coefficients by Review Condition and Testing Mode

Variable	Fixed-Item		Adaptive		Self-Adapted		Total	
	Before	After	Before	After	Before	After	Before	After
Individual Difference								
Test Anxiety[a]	-0.62	-0.63	-0.54	-0.52	-0.15	-0.17	-0.38	-0.41
Criterion								
Vocabulary Criterion Test	0.85	0.83	0.80	0.81	0.85	0.82	0.84	0.83
Verbal Self-Concept[b]	0.44	0.46	0.65	0.63	0.23	0.21	0.46	0.45
Self-Reported College GPA[c]	0.27	0.27	0.64	0.66	0.51	0.53	0.46	0.47

[a]The difference between the fixed-item and self-adapted correlations was significant both before (Z= -2.20, p < .05) and after review (Z= -2.19, p < .05).

[b]The difference between the adaptive and self-adapted correlation was significant both before (Z= 2.06, p < .05) and after review (Z= 2.01, p < .05).

[c]The difference between the adaptive and fixed-item correlation was significant after review (Z= 1.98, p < .05).

Figure 1
Testing Mode by Anxiety Interaction for Ability Scores

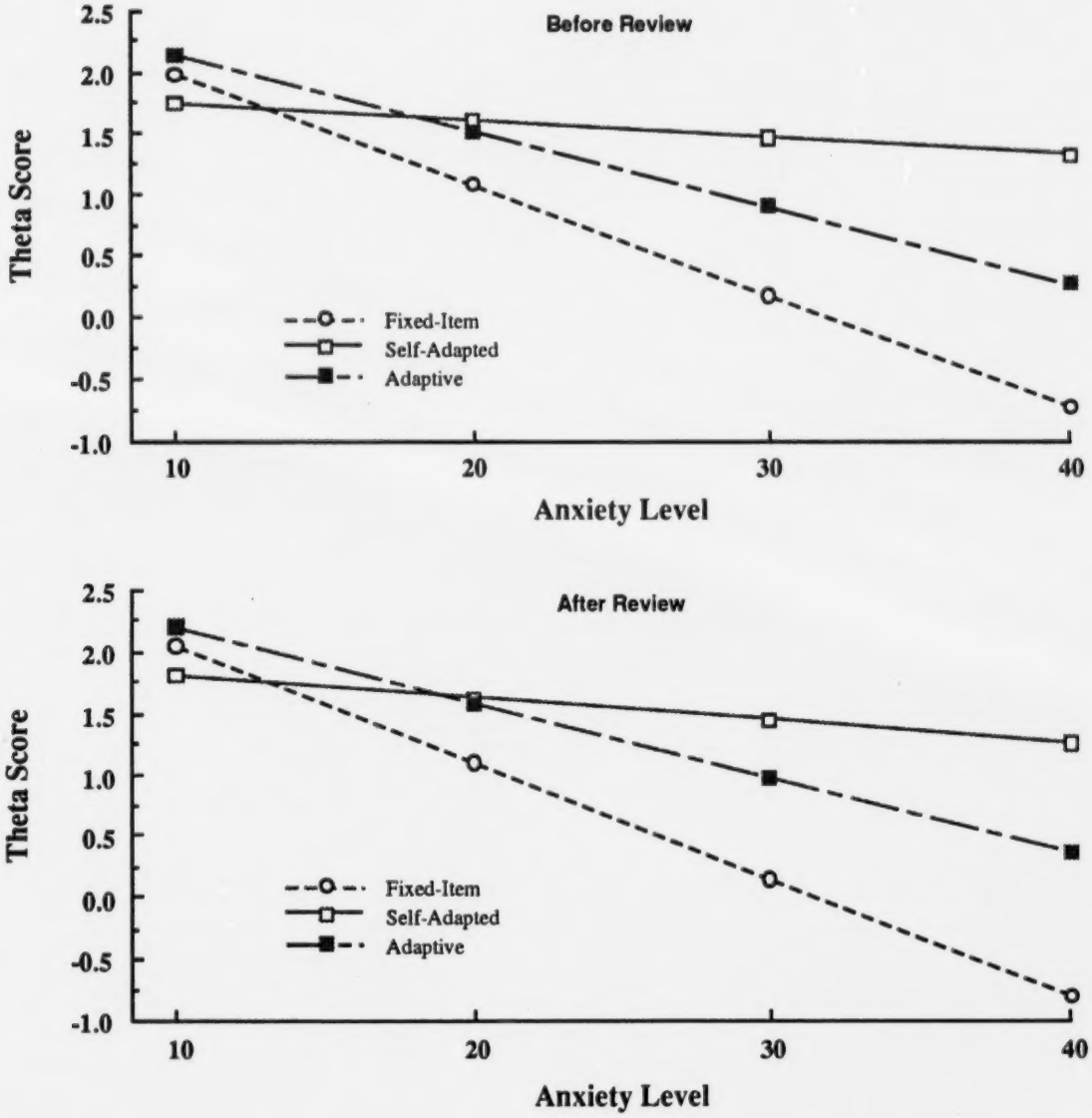

Figure 2
Review by Testing Mode Interaction for Test Administration Time

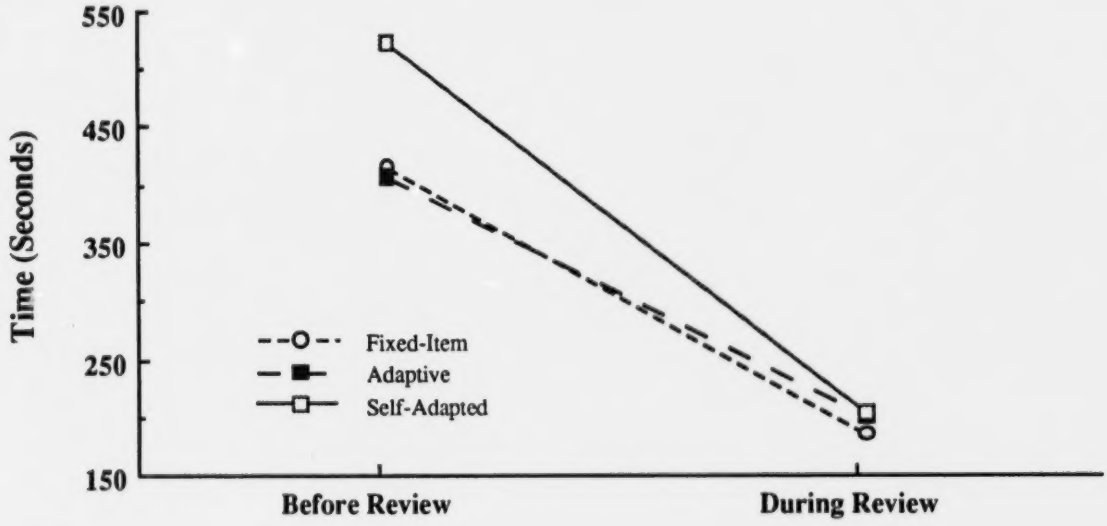

Figure 3
Relationship Between Test Anxiety and Administration Time Before and During Review